

# Synthetic Tools for Molecular Biology

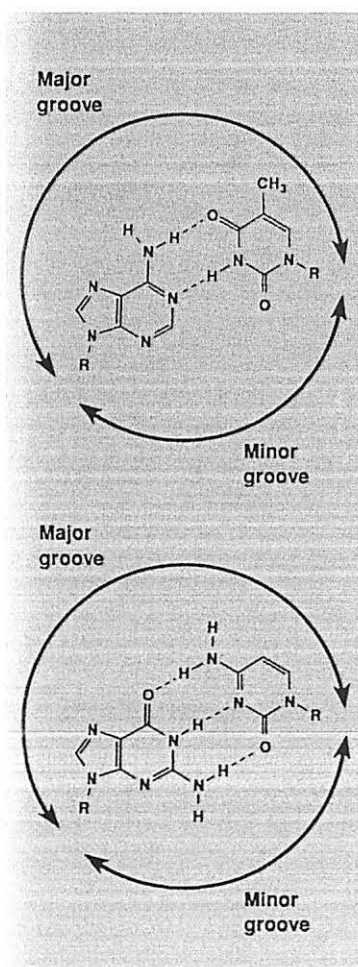
Peter B. Dervan

Chemistry has made tremendous advances over the past four decades in the broad fields of synthesis and understanding chemical reactivity. In that same time span, a series of revolutionary events occurred in biology. First came the discovery of the double helical structure of DNA in the 1950s by Watson and Crick. This discovery allowed the elucidation of the mechanisms of DNA replication—how DNA makes copies of itself—and DNA transcription and translation—the processes that allow the genetic code to be read and translated into proteins. In the 1970s, the techniques that permit DNA to be cut and spliced in controlled and well-defined ways were invented and the technology of recombinant DNA was born.

Chemistry plays a pivotal role in this biological revolution because biological events involve molecules and molecular interactions. No scientists are better qualified to tackle the problem of determining the structure and the shape of molecules than chemists. Our society is entering an era in which chemists have the opportunity to solve some of the most important problems in biology and biomedical science.

Each strand of DNA consists of a linear polymer of nucleotides, which are phosphorylated versions of four different bases: adenine (A), guanine (G), cytosine (C), and thymidine (T). Physical chemical concepts readily explain what holds these nucleotides to complementary nucleotides on the second strand of the DNA molecules. Hydrogen bonding dictates that the base adenine matches up with the base thymidine and the base guanine matches up with the base cytosine (*see* structures). These basic principles of chemistry make it possible for scientists to explain, in a very fundamental way, how life works.

The quantity of DNA that encodes the entire human genome consists of about three billion such base pairs.



That is three billion chemical bits of information. These chromosomes—humans have 23 pairs—are the information repository, or encyclopedia, for the construction of a human being. Molecular biologists estimate that there are 100,000 to 300,000 segments of information or genes contained on these 23 pairs of chromosomes.

Each gene, typically, is the blueprint for one protein. The three-dimensional shape of proteins gives our bodies form and function and allows the proteins to carry out the complex chemistry of life. Clearly, then, understanding the structure of human chromosomes and genes is one of the first steps toward understanding the fundamental machinery of human existence. We already understand a great deal about those structures, but there is much more to learn. These are enormous molecules, and the subtle and complex interactions among them determine the difference, for example, between health and disease.

During the past few years, there has been considerable discussion about whether scientists should begin the task of first mapping and then sequencing the human genome. This is a task of staggering dimensions. *Physical mapping* means using restriction endonucleases to cut up chromosomes and then determine the order of the resulting fragments. *Genetic mapping*, a somewhat less daunting task that already is being undertaken, means locating on chromosomes about 400 known genetic markers. *Sequencing* is the analysis of the order of each of the three billion base pairs that make up the human genome.

### *Sequencing the Human Genome*

First, researchers will have to separate the 23 human chromosomes and obtain a significant quantity of each one. Once the chromosomes have been sorted, one strategy that has been proposed is to break each chromosome into fragments, each fragment containing about 40,000 base pairs. Each chromosome would yield about 2500 of these fragments, and their order on the chromosomes would then have to be determined. These fragments, in turn, would have to be broken down into fragments of a suitable size for direct sequencing, about 1000 base pairs long. Each chromosome would yield

*The po  
of poss  
of th  
genom*

about 100,000 such fragments. Simple arithmetic shows that, at a rate of about one million bases a day, sequencing the human genome will take 10 years to complete.

The potential applications of possessing the sequence of the entire human genome are enormous. These applications include such things as the ability to gain a basic understanding of why certain individuals are susceptible to genetic disorders like cancer, heart disease, and mental illness.

---

*The potential applications of possessing the sequence of the entire human genome are enormous.*

---

At Caltech, biology professor Leroy Hood and his co-workers have worked on developing the chemistry and instrumentation for sequencing DNA and proteins. Hood would be the first to acknowledge that current automated DNA sequencing is still a relatively immature technology and that it will require significant further development in both the underlying chemistry and the instrumentation before it is ready to take on the task of sequencing the human genome. That work is being done both at Caltech in Hood's group and more recently by a group at DuPont. The step toward sequencing the human genome should be the automation of the entire sequencing procedure, not just the downstream portion of sequencing nucleotides, an effort that may take five to ten years. Within that time, nucleotide sequencing should have improved in speed by a factor of about 100 and improved significantly in accuracy. It should then be ready to be applied to the human genome.

During the 15 to 20 years it will take to map and sequence the human encyclopedia, it is probable that the fields of molecular biology and biological chemistry will make new strides into understanding the mechanisms that control gene expression. Gene expression is the product of an extensive array of interactions among DNA, RNA, and proteins. Chemists want to understand the physical and chemical principles that govern this process. One aspect of this research will be to isolate and determine the structures of protein-DNA assemblies that regulate how genes are expressed. It is conceivable, therefore, that concurrent with the determination of the sequence of human DNA, the "software" of the human genome also will be understood, and synthetic chemists will be able to engineer at the molecular level novel synthetic materials that read unique DNA sequences.

For instance, the natural restriction endonucleases that have been used so powerfully by molecular biologists

in recombinant DNA techniques can recognize a specific DNA sequence, four to eight base pairs in size, and cut the DNA molecule at that site. A consequence of a four-letter alphabet for double helical DNA is that each binding site size yields a specific number of unique sites. There are, for example, 136 unique binding sites consisting of four base pairs and 2080 unique sites consisting of six base pairs. In terms of physical mapping of the genome, this number of unique binding sites limits what might be called the resolution of the map.

### *Improving Restriction Endonucleases*

Natural restriction endonucleases recognize and bind to sequences in DNA from four to eight base pairs long and then cut the DNA at that site. A question synthetic chemists in my laboratory have addressed is whether we can improve on the specificity of natural restriction endonucleases. In other words, is it possible to modify natural DNA-binding molecules or synthesize entirely new one that will recognize unique sequences of 12 to 15 base pairs? Again, because DNA uses a four-letter alphabet, there are 10 million to 100 million such unique DNA sequences, and hence, one could potentially create that many molecules. To be specific at the level of one unique gene in large chromosomal DNA, molecules need to recognize DNA sequences made up of 15 base pairs.

This is a chemistry problem in molecular recognition. Despite all the successes of chemistry in the past 30-40 years, the field of molecular recognition in organic chemistry—how macromolecules fold and fit together—is still in its infancy. The way to solve this problem is not by serendipity or by focusing on one particular sequence, but rather to work toward a general solution. That is, to elaborate a set of chemical principles that govern DNA recognition without trying to anticipate the direct application of the DNA-binding molecule in each case.

However, the eventual applications are important. Once scientists have in hand the physical map of the human genome, and for example, know where the gene that causes a particular genetic disease is located, the question is, What are they going to do about it? If the principles of gene expression and regulation are known, then most likely chemists will be able to build artificial

Figure 1. Affinity molecule combines DNA molecule with a group to

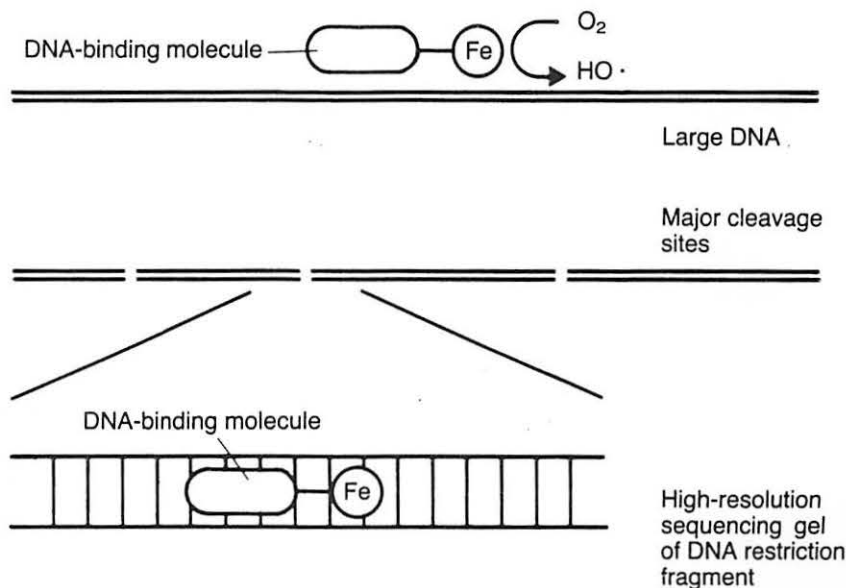
TURE

repressors for specific DNA sequences to control certain disease states.

Because the understanding of the chemistry of molecular complexation is still primitive, the initial efforts in this research borrow heavily from nature. Numerous natural compounds bind to DNA. The restriction endonucleases, as we have seen, bind and cleave DNA. A variety of small molecules, many of them drugs having antibiotic, antiviral, or antitumor activity, also bind to DNA and some also cleave it. Their pharmacological activity results, presumably, from their ability to interfere with some aspect of DNA's function or repair. The goal is to take such small molecules that bind DNA in a modest fashion and improve them to new and novel specificities.

To screen a large number of potential binding sites on DNA, we couple the binding event to the analytical power of gel electrophoresis. This step is accomplished by attaching to the DNA-binding molecule another moiety that cleaves DNA. Thus, we create molecules with two functions: the ability to bind to specific DNA sequences and then the ability to cut the DNA at points adjacent to that sequence. These bifunctional molecules form the basis of a technique called *affinity cleaving*, that is, the binding event on DNA is converted to a sequencing event (see Figure 1). Identification of the preferred binding sites of our designed synthetic molecules is a first step toward addressing the underlying principles of recognition of DNA at the molecular level.

Figure 1. Affinity cleaving technique combines DNA-binding molecule with a group that cuts DNA backbone.



Natural products we have investigated are netropsin and distamycin, di- and tripeptides that are known to bind in the minor groove of right-handed DNA rich in adenine and thymidine bases. The specific interactions involved in this binding process have been determined from the high-resolution crystal structures of a netropsin- and distamycin-oligonucleotide complexes (*see* Figure 2). The question is whether we can improve and alter the specificity of this natural product by building analogs of the molecule that would bind larger DNA fragments. One simple strategy is to build longer versions of the natural product. The longer versions are peptides that consist of four to nine amino acids designed after the structure of the tripeptide. These longer molecules based on distamycin do, in fact, possess a higher specificity for DNA approaching that of restriction endonucleases. So we are moving in the right direction.

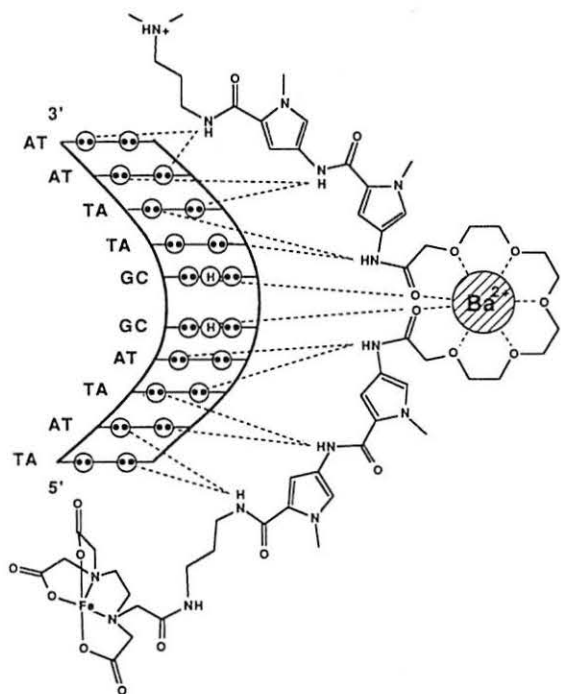


Figure 2. Two netropsin analogs connected by a tetrathylene glycol tether bind a 10 base-pair sequence in the presence of barium metal cation by forming hydrogen bonds (dotted lines) with lone pair electrons •. The EDTA-Fe group is highlighted.

## Molecular Engineering

Another generation of these synthetic molecules will be derived from our understanding of how proteins bind DNA. It might be possible to design synthetic hybrid protein molecules with two structural and functional domains, that is, create synthetic peptides derived from



two different proteins. Researchers in our group have combined 52 amino acids derived from the natural enzyme, Hin-recombinase, which contains 190 amino acids, with three amino acids that constitute the copper binding site in serum albumin. This hybrid synthetic protein possesses two separate structural and functional domains. The 52-amino acid fragment provides the DNA binding specificity of the Hin protein and the three-amino acid fragment binds copper and provides oxidative cleavage of DNA in the presence of hydrogen peroxide and ascorbate.

We have also constructed oligonucleotides to which is attached an EDTA-iron DNA-cleaving function. We have shown that certain such oligonucleotides (15 bases long) will form a triple helix in the major groove of double helical DNA (Figure 3) 11 to 15 base pairs long. This finding makes possible a general solution to the problem of sequence-specific DNA recognition. Within certain constraints, we can now build molecules that can recognize unique specific stretches of DNA 15 base pairs long, which is the level of specificity we need to recognize individual genes in the human genome.

nnected by a  
se-pair sequence  
on by forming  
e pair electrons  
lighted.

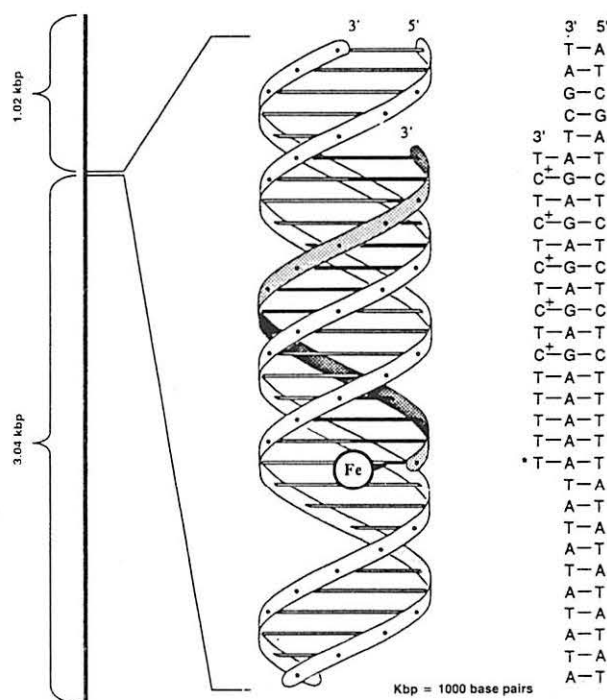


Figure 3. Oligonucleotide recognizes 15-base-pair sequence in 4000-base-pair DNA fragment by forming a triple helix with double-stranded DNA.

### *Catalytic Antibodies*

Turning from the problem of sequence-specific DNA recognition, let me focus briefly on protein recognition and monoclonal antibodies, another important leg of biotechnology. Antibodies are the protein molecules produced by an organism's immune system that bind to foreign molecules in the body and tag them as foreign so that they can be eliminated. In immunological terms, the foreign molecule is an antigen. In an intact organism, a range of subtly different antibodies can be generated for any given antigen; there are, in other words, a number of ways to solve the problem of binding a given antigen. This condition is called a polyclonal response. Antibody-producing B lymphocytes can be fused with a cultured myeloma cell to produce a hybridoma that secretes a specific monoclonal antibody (1).

Recently, independent research performed at the Research Institute of Scripps Clinic (2) and at the University of California, Berkeley (3), shows that, through careful design of an antigen, one can produce an antibody that catalyzes a chemical reaction. Schultz at Berkeley and Lerner and Tramontano at Scripps produced antibodies that catalyze, with a high degree of specificity, hydrolysis of esters and carbonates. The antigenic molecules used to elicit these antibodies are phosphonates, and they mimic the transition state of this hydrolysis reaction.

These researchers are exploring ways to adapt this idea to devise molecules that will cleave protein molecules with great specificity. One direction the researchers are pursuing to introduce catalytic activity into antibodies is production of what might be called *semisynthetic catalytic antibodies*. The idea is to produce an antibody that binds a specific molecule, and chemically to attach catalytic groups such as a metal ion to the antibody to carry out a reaction such as hydrolysis of the peptide bond that links amino acids together in a protein. In such a molecule, the antibody provides the binding specificity and synthetic catalyst provides the needed chemistry.



*The Future*

In short, when the chemical composition of the hardware and software of the human cell is described at the molecular level, the potential will exist to synthesize molecules to control disease in a very precise way. Over the next 20 years, chemists will participate with biologists in solving these extremely important and challenging problems. The solutions will have tremendous practical implications for humanity.

*Acknowledgments*

The author is grateful to the National Institutes of Health, the American Cancer Society, the DARPA University Initiative Research Program, Allied Signal Corporation, Merck Sharp & Dohme, Research Laboratories, Burroughs-Wellcome Company and the Ralph M. Parsons Foundation for generous support. In addition, stimulating and helpful discussions with Dr. Ralph Hirschmann are gratefully acknowledged.

*References*

1. Milstein, C.; Köhler, G. *Nature*, 1975, 257, 495.
2. Tramontano, A.; Janda, K. D.; Lerner, R. A. *Science*, 1986, 234, 1566.
3. Pollack, S. J.; Jacobs, J. W.; Schultz, P. G. *Science*, 1986, 234, 1570.